**ASSIGNMENT 7.2 SCAFFOLD MINNING**

*Data:* Fluorescence Cell-Free Homogeneous Counter Screen to Identify Inhibitors of GFP Chromophore Formation (AID:434968).

*Number of active compounds:* 1764

*Number of inactive compounds:* 349

*Command line in Strip-it:*

* active scaffold mining:

strip-it --input 3237684569133504979.sdf.gz --output active.txt --scaffold scaffold.txt

* inactive scaffold mining:

strip-it --input 176650723242912717.sdf.gz --output inactiv.txt --scaffold scaffold.txt

*Analysis:* using simple python script to calculate the frequency of mined scaffolds in active and inactive compounds. The script is also used to find out the special scaffold which is found in active but not found in inactive.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Top 10 most frequent scaffold in ACTIVE compound data file:

|  |  |
| --- | --- |
| Scaffold in SMILES | Frequency |
| C1CCC(CC1)C1CCCC1 | 75 |
| C1CCC(CC1)CC1CCCC1 | 62 |
| C1CCCCC1 | 40 |
| C1CCC(CC1)CCC1CCCC2C1CCCC2 | 36 |
| C1CCC2C(C1)CCCC2 | 35 |
| C1CCC(CC1)CC1CCCCC1 | 35 |
| C1CCC(CC1)CC1CCC(C1)C1CCCCC | 33 |
| C1CCC(CC1)C1CCC(C1)C1CCCCC1 | 33 |
| C1CCC2C(C1)CCC2 | 32 |
| C1CCC(CC1)C(C1CCCC1)C1CCCC1 | 31 |

1. Top 10 most frequent scaffold in inactive compounds data file:

|  |  |
| --- | --- |
| Scaffold in SMILES | Frequency |
| C1CCC(CC1)CCC1CCCC1' | 10 |
| C1CCC(CC1)CC1CCCCC1' | 9 |
| C1CCC(CC1)CC1CCCC2C1CCCC2' | 9 |
| C(CC1CCCCC1)CC1CCCCC1' | 7 |
| C1CCC(CC1)C1CCCC1' | 6 |
| C(CCC1CCCC1)CCC1CCC(CC1)CCC1CCCCC1 | 5 |
| C1CCC2C(C1)CCC2 | 5 |
| C1CCC(CC1)CCC1CCCCC1' | 5 |
| C1CCCCC1 | 4 |
| C1CCCC1 | 4 |

1. Top 10 most frequent scaffold found BOTH in active and inactive:

|  |  |
| --- | --- |
| Found BOTH in active and inactive | Frequency |
| C1CCC(CC1)C1CCCC1 | 75 |
| C1CCC(CC1)CC1CCCC1 | 62 |
| C1CCCCC1 | 40 |
| C1CCC(CC1)CC1CCCCC1 | 35 |
| C1CCC2C(C1)CCCC2 | 35 |
| C1CCC(CC1)CC1CCC(C1)C1CCCCC1 | 33 |
| C1CCC(CC1)C1CCC(C1)C1CCCCC1 | 33 |
| C1CCC2C(C1)CCC2 | 32 |
| C1CCC(CC1)C(C1CCCC1)C1CCCC1 | 31 |
| C1CCC(CC1)CCC1CCCC1 | 29 |

The frequency here means how many times the scaffold is found in active and it is also found in inactive. When compare this table to the fist table (scaffold in active), it is shown that almost the scaffold and frequency are the same in the two tables except one scaffold: C1CCC(CC1)CCC1CCCC2C1CCCC2 (frequency in active is 36). As taking the fourth table (scaffold in active but not in inactive, listed below) as reference, it conveys the message that this scaffold, C1CCC(CC1)CCC1CCCC2C1CCCC2, is important in the application of drug discovery since it proves that this scaffold distinguishes the compounds between active and inactive.

1. Top 10 most frequent scaffold found in ACTIVE and NOT found in INACTIVE:

|  |  |
| --- | --- |
| Found in ACTIVE and not found in INACTIVE | Frequency |
| C1CCC(CC1)CCC1CCCC2C1CCCC2 | 36 |
| C1CCC(CC1)CC1CCC(C1)CC1CCCCC1 | 15 |
| C1CCC(CC1)C1CCCC2C1C(CC2)C1CCCCC1 | 13 |
| C1CCC(CC1)C(C1CCC(C1)C1CCCCC1)C1CCC(C1)C1CCCCC1 | 13 |
| C1CCC(CC1)C1CCC2C1C(CCC2)C1CCCC1 | 13 |
| C1CCC(CC1)C1CCCC2C1CCC2C1CCCCC1 | 12 |
| C1CC(CCC2CCCCC2)CC(C1)CCC1CCCCC1 | 12 |
| C1CCC(CC1)CC1CCCC(C1)C1CCCCC1 | 10 |
| C1CCC(CC1)CCC(C1CCCCC1)CC1CCCCC1 | 9 |
| C1CCC(CC1)CC1CC(CCC2CCCC2)C2C(C1)CCCC2 | 9 |

The frequency here means that the specific scaffold is found that many times in active and not found once in inactive. These are the special scaffolds that we may pay attention to for drug discovery because these scaffolds make the compounds to be active rather than inactive with the targets. We can make use of these scaffolds to investigate the feature/critical structure in active compounds.